Pattern Recognition – Curve Fitting

Ming-Ju Li

January 26, 2018

Abstract

Find the optimal curve for curve fitting problem by minimizing energy, with or without regularization, Maximum Likelihood Estimation, and Maximum Posterior Estimation under different circumstances.

Contents

1 Introduction

This project is mainly about curve fitting under 4 different methods. These methods are energy minimization, energy minimization with regularization, Maximum likelihood estimation, and Maximum a posterior estimation.

1.1 Data

The data set $\mathbf{X}$, composed of x_n, for n=1,...,50, in this project is distributed uniformly in range [0,1], with the output $\mathbf{t}$, composed of t_n, for n=1,....50, a sinusoid wave with noise distributed in Gaussian distribution.

2 Energy Minimization

The first part of the project is to compute the energy minimization by minimizing the error function.

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1}$$

2.1 Mathematic Derivation

First we create N=50 data, let $\mathbf{w} \equiv (w_0, ..., w_n)^T$, the corresponding observations of these data are $\mathbf{t} \equiv (t_1, ..., t_n)^T$, and we have M-th degree polynomial as our curve fitting model,

$$\mathbf{X} = \begin{bmatrix} x_1^0 & x_1^1 & \dots & x_1^M \\ x_2^0 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & \dots & x_N^M \end{bmatrix} \tag{2}$$

where the energy function given in this is

For the simplicity of deduction, we treat (1) as matrix operation, i.e.,

$$\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 = (\mathbf{Y_n} - \mathbf{t})^2 \tag{3}$$

,where $\mathbf{Y_n} = \mathbf{X} * \mathbf{w}$, while $\mathbf{X}$ is the matrix representation of x_n and $\mathbf{t}$ is the vector representation of t_n. We then can express (1) into

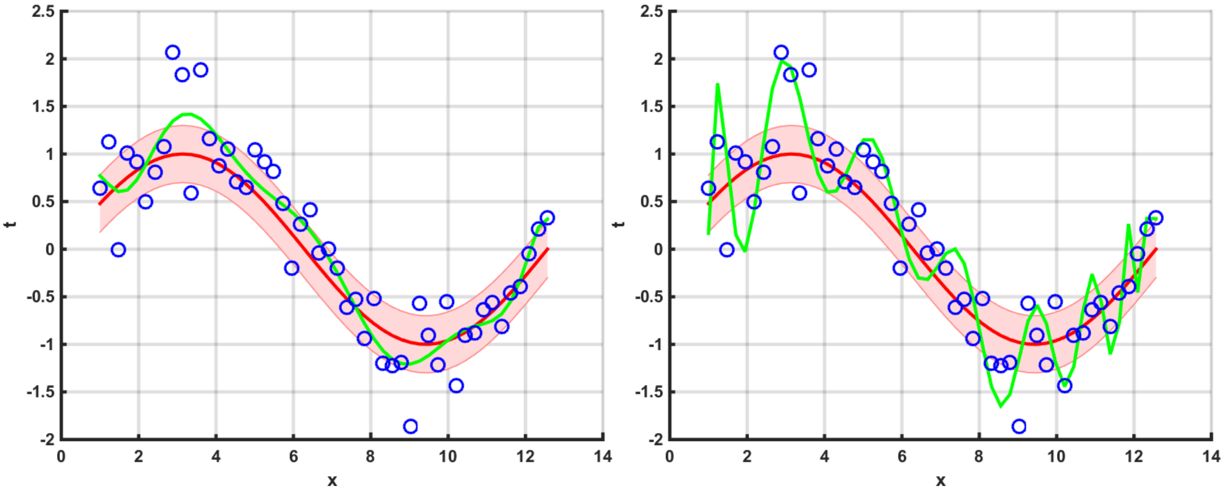
$$E(\mathbf{w}) = \frac{1}{2}[(\mathbf{Xw} - \mathbf{t})^T(\mathbf{Xw} - \mathbf{t})] \tag{4}$$

To minimize (4), we take the partial derivative with respect to $\mathbf{w}$, which will get

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial[(\mathbf{Xw} - \mathbf{t})^T(\mathbf{Xw} - \mathbf{t})]}{\partial \mathbf{w}} \tag{5}$$

$$= \mathbf{X}^T(\mathbf{Xw} - \mathbf{t}) \tag{6}$$

$$= \mathbf{X}^T\mathbf{Xw} - \mathbf{X}^T\mathbf{t} \tag{7}$$



(a) The model with polynomial degree of 10

(b) The model with polynomial degree of 40

Figure 1: Curve fitting by minimizing the error function

To get the optimal solution of $\mathbf{w}$, we set (5) to be zero

$$\mathbf{X}^T\mathbf{X}\mathbf{w}^\star - \mathbf{X}^T\mathbf{t} = 0 \tag{8}$$

we then solve for $\mathbf{w}^\star$,which will yield to

$$\mathbf{w}^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \tag{9}$$

2.2 Result

See figure (1a) and (1b)

3 Energy Minimization with Regularization

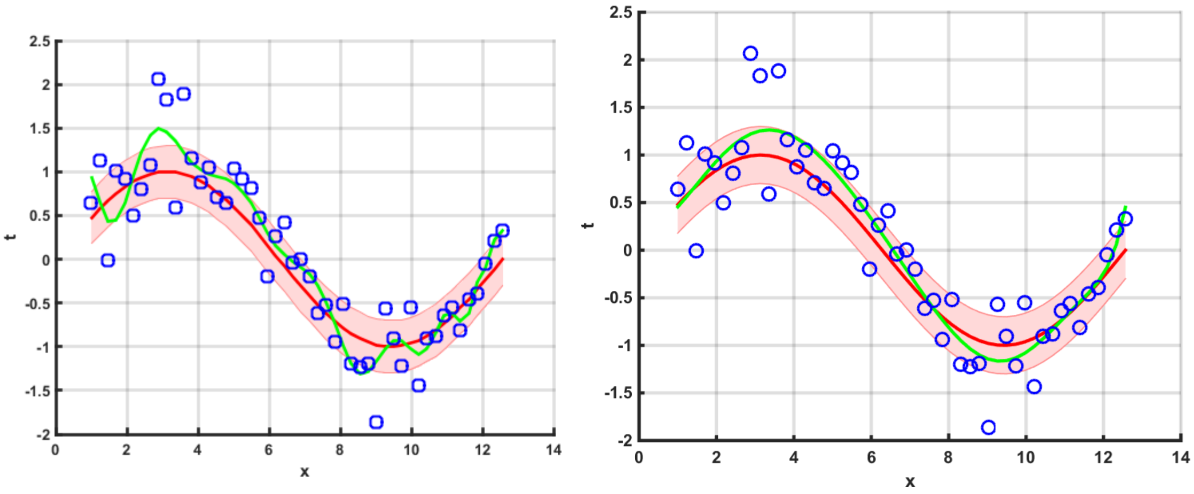
The second part of the project is to compute the energy minimization with regularization coefficient λ, which modified the energy function into

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \tag{10}$$

3.1 Mathematic Derivation

We can also write (10) into matrix and vector representation,

$$\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2 = (\mathbf{Y_n} - \mathbf{t})^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \tag{11}$$



(a) The model of order 40 with coefficient $\ln(\lambda) = -10$

(b) The model of order 40 with coefficient $\ln(\lambda) = 0$

Figure 2: Curve fitting by minimizing error function with regularization

To get the optimal estimation of **w**, we need to minimize (11),

$$\begin{aligned} \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} =& \frac{\partial[(\mathbf{Y_n} - \mathbf{t})^2 + \frac{\lambda}{2}||\mathbf{w}||^2]}{\partial \mathbf{w}} && (12) \\ =& \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) + \lambda\mathbf{w} && (13) \\ =& \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{t} + \lambda\mathbf{w} && (14) \end{aligned}$$

Set (37) equals to zero to get the optimal solution of **w**

$$\mathbf{X}^T\mathbf{X}\mathbf{w}^\star - \mathbf{X}^T\mathbf{t} + \lambda\mathbf{w}^\star = 0 \tag{15}$$

we then solve for $\mathbf{w}^\star$,

$$\mathbf{w}^\star = (\lambda I + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \tag{16}$$

3.2 Result

See figure(2a) and (2b)

4 Maximum Likelihood Estimation

We now use different way to obtain the desired coefficient, **w**. Instead of minimizing the error function mentioned in last two sectors, we now calculate the probability of **w** given a certain data point. Recall that in the introduction, we mentioned that the observation given a certain data has a Gaussian noise.

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x_n, \mathbf{w}), \beta^{-1}) = (\frac{\beta}{2\pi})^{\frac{1}{2}} exp\{-\frac{\beta(t - x_n)^2}{2}\} \tag{17}$$

,where β is the inverse of the variance of the distribution.

We therefore assume that $\mathbf{w}$ also have Gaussian distribution. So the probability of the data set $\mathbf{X}, \mathbf{t}$ will become

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{t}|y(x, \mathbf{w}), \beta^{-1}) = (\frac{\beta}{2\pi})^{\frac{1}{2}} exp\{-\frac{\beta(\mathbf{t}-\mathbf{X})^2}{2}\} \tag{18}$$

For computational convenience, we take logarithm to the likelihood function

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - \mathbf{t})^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \tag{19}$$

Also, scaling the log likelihood by a positive constant does not alter the location of the maximum with respect to $\mathbf{w}$, and so we can ignore the effect of β. Thus, maximizing the log-likelihood function is the equivalent to minimize the negative log-likelihood function.

4.1 Mathematic Derivation

First, we let the likelihood function be

$$L(\beta) = \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \tag{20}$$

then take partial derivative with respect to both β and $\mathbf{w}$,

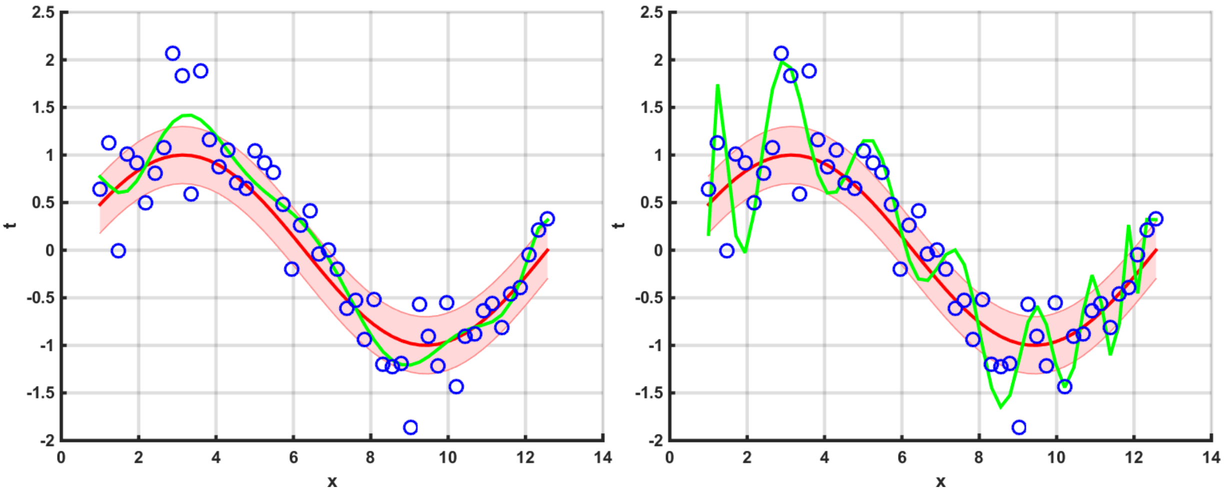
$$\begin{aligned}
\frac{\partial L(\beta)}{\partial \mathbf{w}} =& \frac{\partial[(\mathbf{X}\mathbf{w}-\mathbf{t})^T(\mathbf{X}\mathbf{w}-\mathbf{t}) + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)]}{\partial \mathbf{w}} && (21)\\
=& \mathbf{X}^T(\mathbf{X}\mathbf{w}-\mathbf{t}) && (22)\\
=& \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{t} && (23)\\
& && (24)
\end{aligned}$$

Recall that we have ignored the effect of β on $\mathbf{w}$. Therefore, we only need to consider the the likelihood funcion $y(x_n, \mathbf{w}) - \mathbf{t}$, which can be written as $[(\mathbf{X}\mathbf{w}-\mathbf{t})^T(\mathbf{X}\mathbf{w}-\mathbf{t})]$. Solving $\mathbf{w}$ will get the same result as 10

$$\mathbf{w}^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \tag{25}$$

We have the optimal solution $\mathbf{w}^\star$, now we want to get the optimal β in order to fully describe the Gaussian distribution in (18). We then take the derivative to the likelihood function (20) with respect to β. Note that we cannot ignore the effect of β on the first term at this point since we are trying to find the optimal value of β.

$$\begin{aligned}
\frac{\partial L(\beta)}{\partial \beta} =& \frac{\partial[-\frac{\beta}{2}(\mathbf{X}\mathbf{w}-\mathbf{t})^T(\mathbf{X}\mathbf{w}-\mathbf{t}) + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)]}{\partial \mathbf{w}} && (26)\\
=& -(\mathbf{X}\mathbf{w}-\mathbf{t})^T(\mathbf{X}\mathbf{w}-\mathbf{t}) + \frac{N}{\beta} && (27)
\end{aligned}$$



(a) The model with polynomial degree of 10 (b) The model with polynomial degree of 40

Figure 3: Curve fitting by MLE with polynomial degree of 40

By setting the equation (26) to zero, we can get the optimal value of β.

$$-(\mathbf{Xw}-\mathbf{t})^T(\mathbf{Xw}-\mathbf{t}) + \frac{N}{\beta} = 0 \tag{28}$$

$$\Rightarrow \beta_{ML} = \frac{N}{(\mathbf{Xw}-\mathbf{t})^T(\mathbf{Xw}-\mathbf{t})} \tag{29}$$

or

$$\frac{1}{\beta_{ML}} = \frac{(\mathbf{Xw}-\mathbf{t})^T(\mathbf{Xw}-\mathbf{t})}{N} \tag{30}$$

4.2 Result

see Figure(3a) and Figure(3b)

5 Maximum a Posterior Estimation

In the previous section, we have the maximum likelihood estimation for $\mathbf{w}$ given a certain data point. But there is one thing we need to consider. Recall the Bayesian rule

$$Posterior \propto Liklihood \times Prior \tag{31}$$

To get the posterior estimation function, we need to multiply the likelihood we estimated in the previous section (see equation (18)) with the prior function, which is normally come from the empirical statistic or some observations. We can write the posterior function as

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \times p(\mathbf{w}|\alpha) \tag{32}$$

,where

$$p(\mathbf{w}|\alpha) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = (\frac{\alpha}{2\pi})^{\frac{M+1}{2}} exp - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{33}$$

5.1 Mathematic Derivation

By multiplying eq(18) and eq(33) we can get the posterior function $p(\mathbf{w}|\mathbf{X},\mathbf{t},\alpha,\beta)$. We can now determine $\mathbf{w}$ by finding the most probable value of $\mathbf{w}$ given the data, in other words by maximizing the posterior function.

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(\mathbf{X},\mathbf{w})-\mathbf{t}\}^2+\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{34}$$

Maximizing the posterior function is equivalent to minimizing the regularized error function encountered in form(10), with a regularization parameter $\lambda = \alpha/\beta$.
Again, the positive constantβ does not alter the location of the maximum with respect to $\mathbf{w}$, we could ignore it while evaluating the MAP.

$$\begin{aligned}\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} =& \frac{\partial[(\mathbf{Y_n}-\mathbf{t})^2+\frac{\alpha}{2}||\mathbf{w}||^2]}{\partial \mathbf{w}} && (35)\\ =&\mathbf{X}^T(\mathbf{X}\mathbf{w}-\mathbf{t})+\alpha\mathbf{w} && (36)\\ =&\mathbf{X}^T\mathbf{X}\mathbf{w}-\mathbf{X}^T\mathbf{t}+\alpha\mathbf{w} && (37)\end{aligned}$$

Set (37) equals to zero to get the optimal solution of $\mathbf{w}$

$$\mathbf{X}^T\mathbf{X}\mathbf{w}^\star-\mathbf{X}^T\mathbf{t}+\alpha\mathbf{w}^\star=0 \tag{38}$$

we then solve for $\mathbf{w}^\star$,

$$\mathbf{w}^\star=(\alpha I+\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \tag{39}$$

5.2 Result

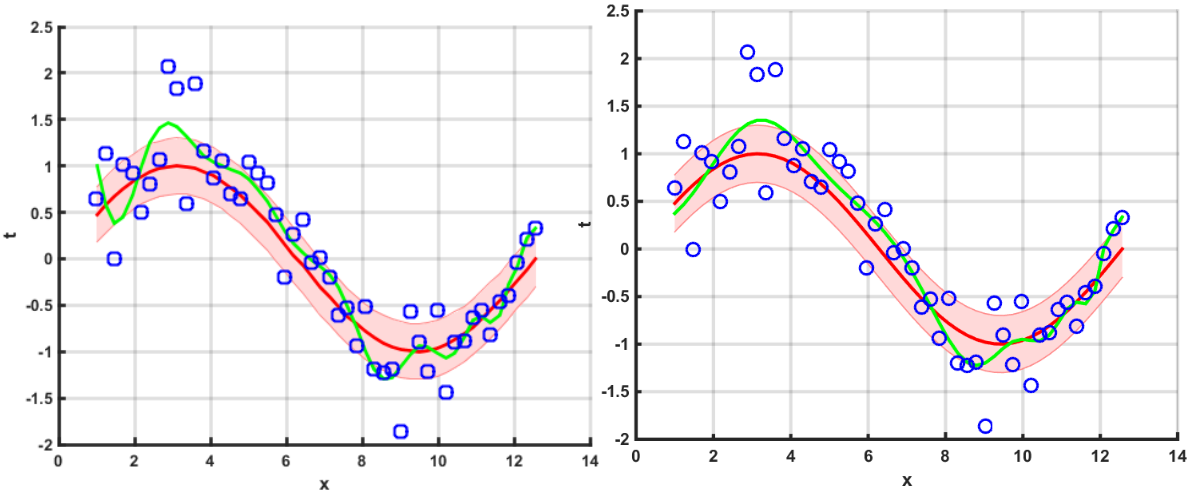
See figure(4a) and (4b)

6 Conclusion

In this project, the energy minimization is used in the first part to get a solution that will optimize the error function (equation(1)). The result seems satisfying . However, another problem will easily appear, *over-fitting*, since there exists no limitation of $\mathbf{w}$. Therefore, the second method, energy minimization with regularization, is applied to deal with this problem (equation(10)).

With the additional parameter λ, it is possible to limit the deal with over-fitting problem (equation(16)). Manipulating λ can get a more desirable curve. This method can do the same thing as the first method does, but with a satisfying result. But if the amount of the training data is not enough to establish the model, *Maximum Likelihood Estimation*, or *MLE* is employed.

Instead of finding the polynomial that minimize the error function, this method finds the distribution of the observations $\mathbf{t}$ given the training data, ,the coefficient of a curve, and a variance (equation(17)). Now, we are trying to find the polynomial coefficient $\mathbf{w}$ that maximize the likelihood of $\mathbf{t}$, $p(t|x,\mathbf{w},\beta)$. Nonetheless, it is unfair to merely employ



(a) The model of order 40 with coefficient $\alpha = 2\exp(-10)$

(b) The model of order 40 with coefficient $\alpha = 2\exp(0)$

Figure 4: Curve fitting using MAP under different α

the likelihood function and ignore the importance of the prior, which evokes the idea of *Maximum Posterior Estimation*, or *MAP*.

By employing the Bayes' rule, it is more reasonable to describe the probability distribution (equation (32)). The priors are usually come from the results of experiments or the statistical tendency, which could possibly make the distribution more objective instead of merely estimate using the likelihood function. It also puts some constraints to the distribution function by introducing the hyperparameters. Same as MLE, MAP evaluates the coefficients **w** by the posterior probability distribution instead of minimizing the error functions (equation (34)).